UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/813,642	03/30/2004	Ara V. Nefian	884.C05US1	4943

21186 7590 10/28/2008

SCHWEGMAN, LUNDBERG & WOESSNER, P.A.
P.O. BOX 2938
MINNEAPOLIS, MN 55402

EXAMINER	
YEN, ERIC L	
ART UNIT	PAPER NUMBER
2626	
MAIL DATE	DELIVERY MODE
10/28/2008	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/813,642	NEFIAN ET AL.	
	Examiner	Art Unit	
	ERIC YEN	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *21 July 2008*.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-5,7-10 and 12-27* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-5, 7-10, 12-27* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

Response to Amendment

1. In response to the Office Action mailed 4/21/08, applicant has submitted an amendment filed 7/21/08.

Claims 1, 8, 15, 20, and 25, have been amended.

Response to Arguments

2. Applicant's arguments with respect to claims 1, 8, 15, 20, and 25, have been considered but are moot in view of the new ground(s) of rejection.

Applicant argues that “none of the references discuss capturing the video and audio together and at different rates from one another and then separating them from one another and re-associating based on time when captured" (Amendment, page 9).

While the capture at different rates is not explicitly taught by Kimura, it is not clear how Katsumi doesn’t teach the portion of the amended limitations that covers capturing, separating, and re-associating based on time. This is at least because of the broad claim language amended.

For example in Claim 1, the claim language recites “wherein the visual features are separated from the audio and processed separately... and separated based on time when each was captured" (lines 7-10). In Katsumi's Figure 3, the system receives audio information and video information from the line interface which among other things "converts the transmission signals received... into video signals and audio signals

and supplies these signals to the video processing portion and the audio processing portion, respectively" (col. 4, lines 34-46). This portion alone at least suggests that what is received form the conference terminals is either a multiplexed or other combined form of the audio and video received at the terminal, and later separated from transmission format to audio only and video only signals that go into the corresponding audio/video processing portions.

The claims also recite "based on time" (Claim 1, lines 9-10) or limitations similar in scope. Applicant does not claim that the association is based on, for example, a time stamp as described in the Specification, nor does applicant describe that the "association" is any sort of re-combining. All that is required from the claim is that there is some association from the time when it is received. This limitation is met from the description in Kimura that "the 'speaking attendee determination... based on audio signal'... and... based on video signal'" (col. 6, lines 57-67), where the audio signal and video signal in this portion are obviously from the same time since it would not make sense to delay either signal and risk making a determination of speaking at a time when the person at the conference terminal is not speaking. Therefore, the audio signal and the video signal are associated with one another as portions of audio/video used to determine whether someone is actually speaking or not at a specific time period, which is a function of the time of capture since it would be illogical to capture audio at one time, and the video sometime later, to make a determination of whether someone is speaking at one time using both signals that exist at two entirely different times when a

user may or may not be speaking at the two different times. This applies, also, to Claim 8 since they are compared based on the time they exist.

Claim Rejections - 35 USC § 103

3. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

1. **Claims 1-5, 8-10, and 12-27** are rejected under 35 U.S.C. 103(a) as being unpatentable over Katsumi Patent No.: US 6,369,846 ("KATSUMI") in view of Nefian Pub. No.: US 2003/0212557 ("NEFIAN") and Lubiarz et al. (US 7,003,452), hereafter Lubiarz, and Holmes et al. (US 5,506,932), hereafter Holmes.

2. Regarding **claim 1**, KATSUMI teaches a method, comprising:

electronically capturing visual features associated with a speaker speaking ("speaking attendee determination information based on video signal", KATSUMI, column 6, lines 52-53);

electronically capturing audio and wherein the visual features are separated form the audio and processed separately, and separated based on time when each was captured ("speaking attendee determination information based on audio signal",

KATSUMI, column 6, lines 50-51; col. 6, lines 57-67; col. 4, lines 34-46; Figure 3; See Response to Arguments);

matching selective portions of the audio with the visual features ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63); and

identifying the remaining and unmatched portions of the audio as potential noise not associated with the speaker speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

However, KATSUMI does not disclose that the method occurs during a training session; that the visual features include a face recognition of the speaker and a mouth recognition within pixels associated with the face that detects when the mouth is moving and when the mouth is not moving by differences in the pixels from frame to frame in the captured visual features during the training session; and detecting frequencies in the captured audio during same time slices of the training session for the captured visual features when the mouth is detected as moving and when the mouth is detected as not moving.

In the same field of audiovisual processing, NEFIAN teaches:

a training session ("training network and speech recognition module 18", NEFIAN, paragraph [0012]);

visual features including a face recognition of the speaker and a mouth recognition within pixels associated with the face that detects when the mouth is moving and when the mouth is not moving by differences in the pixels from frame to frame in the captured visual features during the training session (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions); and

detecting frequencies in the captured audio ("13 MFCC coefficients extracted from a window of 20 ms", NEFIAN, paragraph [0055]) during same time slices of the training session for the captured visual features when the mouth is detected as moving and when the mouth is detected as not moving ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

Katsumi, in view of Nefian, fail to teach where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, and where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session.

Lubiarz teaches where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, and where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session (“speech recognition”, col. 2, lines 28-37; “detects silence… detects the presence of voice activity… optimized… for each of the frequency bands”, col. 5, lines 17-34).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian, to include the teaching of Lubiarz of where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, and where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, in order to ensure that the most effective processing for a particular type of signal is applied to that type of signal, as described by Lubiarz (col. 1, lines 5-10).

Katsumi, in view of Nefian and Lubiarz, fail to teach wherein the visual features and audio are initially captured at a different rate from one another.

Holmes teaches wherein the visual features and audio are initially captured at a different rate from one another (“synchronizes two or more streams of data", col. 1, lines 34-44; "rate of video clock... ratio... adjustable to yield a fixed audio sampling rate", col. 5, lines 33-48;”synch pulses… distinguished by their timing relationships”, col. 4, lines 33-45; where ratios are generally known to be more than just a 1 to 1 ratio, and so the teaching of adjustable ratios implies that the sample rates of the video and audio are not the same; Alternatively, the rates are “different” in nature, since video is, for example,

pixel/image frame rate and audio is audio sampling rate, which means that they are different in nature and not in number. The synch pulses also are an indicator of a time of frame capture because it provides a reference point for comparing when one frame was captured relative to another and where to link the corresponding audio)

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian and Lubiarz, to include the teaching of Holmes of wherein the visual features and audio are initially captured at a different rate from one another in order to prevent either the audio or video from running ahead of the other at any point which could throw off later analysis, as described by Holmes (col. 2, lines 44-64).

3. Regarding **claim 2**, KATSUMI further teaches:

electronically capturing additional visual features associated with a different speaker speaking ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57); and

matching some of the remaining portions of the audio from the potential noise with the additional speaker speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

4. Regarding **claim 3**, NEFIAN further teaches generating parameters associated with the matching and the identifying ("audio processing and visual feature extraction", NEFIAN, paragraph [0012]) and providing the parameters to a Bayesian Network which models the speaker speaking ("video data must be fused with audio data using… a coupled hidden Markov model [HMM]", NEFIAN, paragraph [0023], where the HMM is a dynamic Bayesian network).

5. Regarding **claim 4**, NEFIAN further teaches that electronically capturing the visual features further includes processing a neural network ("neural network", NEFIAN, paragraph [0014]) against electronic video associated with the speaker speaking ("speaker's face in a video sequence", NEFIAN, paragraph [0014]), wherein the neural network is trained to detect and monitor the face of the speaker ("face detection", NEFIAN, paragraph [0014]).

6. Regarding **claim 5**, NEFIAN further teaches filtering the detected face of the speaker to detect movement or lack of movement in the mouth of the speaker ("after the face is detected, mouth region discrimination is usual", NEFIAN, paragraph [0015]).

7. Regarding **claim 8**, KATSUMI teaches a method, comprising:

monitoring an electronic video of a first speaker and a second speaker ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57);

concurrently capturing audio associated with the first and second speaker speaking ("voices may be contained in the audio signal", KATSUMI, column 3, line 1);

analyzing the video to detect when the first and second speakers are moving their respective mouths ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25);

the audio and video are separated and then compared based on time ("speaking attendee determination information based on audio signal", KATSUMI, column 6, lines 50-51; col. 6, lines 57-67; col. 4, lines 34-46; Figure 3; See Response to Arguments); and

matching portions of the captured audio to the first speaker and other portions to the second speaker based on the analysis and wherein at least some points in the training session indicate that the first and second speakers are simultaneously speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

However, KATSUMI does not disclose a training session for face recognition of the first and second speakers; indications as to when mouths for the first and second speakers are moving and not moving from frame to frame of the video during the training session; audio separated from video and matched back to a corresponding portion of the video via a particular time slice associated with both the audio and the video; detecting differences in pixels within the faces occurring from frame to frame of

the video for each of the speakers; and detecting changes in frequencies within the audio for a same time slice that indicates a particular mouth of one of the speakers is moving and by noting a particular frequency for a particular one of the speakers, and discerning what each is saying based on their respective frequencies that were noted.

In the same field of audiovisual processing, NEFIAN teaches:

a training session ("training network and speech recognition module 18", NEFIAN, paragraph [0012]);

indications as to when mouths for the first and second speakers are moving and not moving from frame to frame of the video during the training session (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions);

audio separated from video ("audiovisual data is separately subjected to audio processing and visual feature extraction 14", NEFIAN, paragraph [0012]) and matched back to a corresponding portion of the video ("video data must be fused with audio data", NEFIAN, paragraph [0023]) via a particular time slice associated with both the audio and the video ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]);

detecting differences in pixels within the faces occurring from frame to frame of the video for each of the speakers (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions); and

detecting changes in frequencies within the audio ("13 MFCC coefficients extracted from a window of 20 ms", NEFIAN, paragraph [0055]) for a same time slice

that indicates a particular mouth of one of the speakers is moving ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]) and by noting a particular frequency for a particular one of the speakers ("13 MFCC coefficients extracted from a window of 20 ms", NEFIAN, paragraph [0055]), and discerning what each is saying based on their respective frequencies that were noted ("for audio-only speech recognition", NEFIAN, paragraph [0055]).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

Katsumi, in view of Nefian, fail to teach where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, and where a particular band of frequency is noted to determine when the speaker is speaking and when the speaker is not speaking during the training session.

Lubiarz teaches where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, and where a particular band of frequency is noted to determine when the speaker is speaking and when the speaker is not speaking

during the training session (“speech recognition”, col. 2, lines 28-37; “detects silence… detects the presence of voice activity… optimized… for each of the frequency bands”, col. 5, lines 17-34; exceeding a threshold notes a frequency band).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian, to include the teaching of Lubiarz of where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, and where a particular band of frequency is noted to determine when the speaker is speaking and when the speaker is not speaking during the training session, in order to ensure that the most effective processing for a particular type of signal is applied to that type of signal, as described by Lubiarz (col. 1, lines 5-10).

Katsumi, in view of Nefian and Lubiarz, fail to teach wherein the video and audio are initially captured at a different rate from one another.

Holmes teaches wherein the video and audio are initially captured at a different rate from one another (“synchronizes two or more streams of data", col. 1, lines 34-44; "rate of video clock... ratio... adjustable to yield a fixed audio sampling rate", col. 5, lines 33-48;”synch pulses… distinguished by their timing relationships”, col. 4, lines 33-45; where ratios are generally known to be more than just a 1 to 1 ratio, and so the teaching of adjustable ratios implies that the sample rates of the video and audio are not the same; Alternatively, the rates are “different” in nature, since video is, for example, pixel/image frame rate and audio is audio sampling rate, which means that they are

different in nature and not in number. The synch pulses also are an indicator of a time of frame capture because it provides a reference point for comparing when one frame was captured relative to another and where to link the corresponding audio)

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian and Lubiarz, to include the teaching of Holmes of wherein the video and audio are initially captured at a different rate from one another in order to prevent either the audio or video from running ahead of the other at any point which could throw off later analysis, as described by Holmes (col. 2, lines 44-64).

8. Regarding **claim 9**, NEFIAN further teaches modeling the analysis for subsequent interactions with the first and second speakers ("the result is a model for the underlying process", NEFIAN, paragraph [0033]).

9. Regarding **claim 10**, NEFIAN further teaches that analyzing further includes processing a neural network ("neural network", NEFIAN, paragraph [0014]) for detecting the faces of the first and second speakers ("speaker's face in a video sequence", NEFIAN, paragraph [0014]) and processing vector classifying algorithms to detect when the first and second speakers' respective mouths are moving or not moving (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations is performed on the mouth regions).

10. Regarding **claim 13**, KATSUMI further teaches identifying selective portions of the captured audio as noise if the selective portions have not been matched to the first speaker or the second speaker ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

11. Regarding **claim 14**, NEFIAN further teaches that matching further includes identifying time dependencies associated with when selective portions of the electronic video were monitored and when selective portions of the audio were captured ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]).

12. Regarding **claim 15**, KATSUMI teaches a system, comprising:

a camera (see KATSUMI, column 4, lines 22-23, the conference terminals produce video signals, therefore a camera is inherent);

a microphone (see KATSUMI, column 4, lines 22-23, the conference terminals produce audio signals, therefore a microphone is inherent); and

a processing device ("MCU", KATSUMI, column 4, line 21), wherein the camera captures video of a speaker and communicates the video to the processing device, the

microphone captures audio associated with the speaker and an environment of the speaker and communicates the audio to the processing device ("the conference terminals 6a to 6c multiplex video signals and audio signals of locations [A] to [C]… and transmit the transmission signals to the MCU", KATSUMI, column 4, lines 22-26) and the video and audio separated from one another ("speaking attendee determination information based on audio signal", KATSUMI, column 6, lines 50-51; col. 6, lines 57-67; col. 4, lines 34-46; Figure 3; See Response to Arguments);, the processing device includes instructions that identifies visual features of the video where the speaker is speaking ("speaking attendee determination information based on video signal", KATSUMI, column 6, lines 52-53) and uses time dependencies to match portions of the audio to those visual features ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

However, KATSUMI does not disclose a plurality of frames within a period of time designated as a training session and each frame associated with a particular time slice; audio associated with the particular time slice of the training session; and a processing device that recognizes a face of the speaker in each frame of the video and a mouth within the face and detects when the mouth is moving or not moving from frame to frame of the video by changes in pixels associated with the mouth, and wherein when the mouth is moving a detected change in frequency within the same time slice of the

audio identifies the speaker as speaking and a particular frequency that uniquely identifies the speaker when the speaker is speaking.

In the same field of audiovisual processing, NEFIAN teaches:

a plurality of frames ("digital form including but not limited to MPEG-2", NEFIAN, paragraph [0011]) within a period of time designated as a training session ("training network and speech recognition module 18", NEFIAN, paragraph [0012]) and each frame associated with a particular time slice (MPEG frames are inherently associated with a time slice);

audio associated with the particular time slice of the training session ("video data must be fused with audio data", NEFIAN, paragraph [0023]); and

a processing device that recognizes a face of the speaker in each frame of the video ("speaker's face in a video sequence", NEFIAN, paragraph [0014]) and a mouth within the face and detects when the mouth is moving or not moving from frame to frame of the video by changes in pixels associated with the mouth (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions), and wherein when the mouth is moving a detected frequency within the same time slice of the audio identifies the speaker as speaking ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]; "13 MFCC coefficients extracted from a window of 20 ms", NEFIAN, paragraph [0055]).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN

with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

Katsumi, in view of Nefian, fail to teach where the detecting frequencies is detecting bands of frequencies, determining that the speaker is speaking, and a particular band of frequency that uniquely identifies the speaker when the speaker is not speaking.

Lubiarz suggests where the detecting frequencies is detecting bands of frequencies, determining that the speaker is speaking, and a particular band of frequency that uniquely identifies the speaker when the speaker is not speaking ("speech recognition", col. 2, lines 28-37; "detects silence… detects the presence of voice activity… optimized… for each of the frequency bands", col. 5, lines 17-34; speech occurs over a relatively defined range of frequencies [i.e., band], and so this range is a unique band that identifies whether the speaker is speaking or not speaking. Also, since Lubiarz teaches processing only when speech is present, Lubiarz suggests where there is an indicator [e.g., the voice activity decision] that is output at a particular time to tell the system that it is time to perform the speech recognition with face movement recognition taught by Katsumi and Nefian).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian, to include the teaching of Lubiarz of where the detecting frequencies is detecting bands of frequencies, determining that the speaker is speaking, and a particular band of frequency that uniquely identifies the speaker when the speaker is not speaking, in order to ensure that the most effective

processing for a particular type of signal is applied to that type of signal, as described by Lubiarz (col. 1, lines 5-10).

Katsumi, in view of Nefian and Lubiarz, fail to teach wherein the video and audio are initially captured at a different rate from one another.

Holmes teaches wherein the video and audio are initially captured at a different rate from one another (“synchronizes two or more streams of data", col. 1, lines 34-44; "rate of video clock... ratio... adjustable to yield a fixed audio sampling rate", col. 5, lines 33-48;”synch pulses… distinguished by their timing relationships”, col. 4, lines 33-45; where ratios are generally known to be more than just a 1 to 1 ratio, and so the teaching of adjustable ratios implies that the sample rates of the video and audio are not the same; Alternatively, the rates are “different” in nature, since video is, for example, pixel/image frame rate and audio is audio sampling rate, which means that they are different in nature and not in number. The synch pulses also are an indicator of a time of frame capture because it provides a reference point for comparing when one frame was captured relative to another and where to link the corresponding audio)

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian and Lubiarz, to include the teaching of Holmes of wherein the video and audio are initially captured at a different rate from one another in order to prevent either the audio or video from running ahead of the other at any point which could throw off later analysis, as described by Holmes (col. 2, lines 44-64).

13. Regarding **claim 16**, KATSUMI further teaches that the captured video also includes images of a second speaker ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57) and the audio includes sounds associated with the second speaker ("voices may be contained in the audio signal", KATSUMI, column 3, line 1), and wherein the instructions matches some portions of the audio to the second speaker when some of the visual features indicate the second speaker is speaking ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

Regarding **claim 17**, NEFIAN further teaches instructions that interact with a neural network ("neural network", NEFIAN, paragraph [0014]) to detect the face of the speaker from the captured video ("speaker's face in a video sequence", NEFIAN, paragraph [0014]).

14. Regarding **claim 18**, NEFIAN further teaches that the instructions interact with a pixel vector algorithm to detect when the mouth associated with the face moves or does not move within the captured video (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions).

15. Regarding **claim 19**, NEFIAN further teaches that the instructions generate parameter data ("audio processing and visual feature extraction", NEFIAN, paragraph [0012]) that configures a Bayesian network ("video data must be fused with audio data using… a coupled hidden Markov model [HMM]", NEFIAN, paragraph [0023], where the HMM is a dynamic Bayesian network) which models subsequent interactions with the speaker ("the result is a model for the underlying process", NEFIAN, paragraph [0033]) to determine when the speaker is speaking and to determine appropriate audio to associate with the speaker speaking in the subsequent interactions ("speech recognition", NEFIAN, paragraph [0023]).

16. Regarding **claim 20**, KATSUMI teaches a machine accessible medium having associated instructions, which when accessed, results in a machine performing:

separating audio and video associated with a speaker speaking into separate frames for analysis (see KATSUMI, FIG. 3, the audio processing is separate from the video processing);

identifying visual features from the video that indicate a mouth of the speaker is moving or not moving ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25); and

associating portions of the audio with selective ones of the visual features that indicate the mouth is moving ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination

information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

However, KATSUMI does not disclose: a training session, wherein each frame associated with a same time line to permit specific frames of the video to be matched to specific frequencies of the audio during a same time slice occurring along the time line for the training session; identifying a face of the speaker and then identifying pixels within the face that represents the mouth and then noting changes in the pixels from frame to frame of the video along the time line; and matching frequencies of the audio with detected movements of the mouth during a same time period within the time line and associating a frequency with the speaker when the mouth is moving.

In the same field of audiovisual processing, NEFIAN teaches:

a training session ("training network and speech recognition module 18", NEFIAN, paragraph [0012]), wherein each file associated with a same time line to permit specific frames of the video to be matched to specific frequencies of the audio during a same time slice occurring along the time line for the training session ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]);

identifying a face of the speaker ("speaker's face in a video sequence", NEFIAN, paragraph [0014]) and then identifying pixels within the face that represents the mouth and then noting changes in the pixels from frame to frame of the video along the time line mouth (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions); and

matching frequencies of the audio with detected movements of the mouth during a same time period within the time line ("video data must be fused with audio data", NEFIAN, paragraph [0023]) and associating a particular frequency with the speaker when the mouth is moving ("13 MFCC coefficients extracted from a window of 20 ms", NEFIAN, paragraph [0055]).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

Katsumi, in view of Nefian, fail to teach where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, and where the frequencies are bands of frequencies.

Lubiarz teaches where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, and where the frequencies are bands of frequencies ("speech recognition", col. 2, lines 28-37; "detects silence… detects the presence of voice activity… optimized… for each of the frequency bands", col. 5, lines 17-34; exceeding a threshold indicates speech in a frequency band, and performing the

appropriate processing as per Katsumi and Nefian includes matching the speech portion with the appropriate video).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian, to include the teaching of Lubiarz of where the detecting frequencies is detecting bands of frequencies in the captured audio during time slices, where the detecting bands of frequencies is to determine when the speaker is speaking and when the speaker is not speaking during the training session, and where the frequencies are bands of frequencies, in order to ensure that the most effective processing for a particular type of signal is applied to that type of signal, as described by Lubiarz (col. 1, lines 5-10).

Katsumi, in view of Nefian and Lubiarz, fail to teach the audio and video originally captured at a different rate from one another and later associated via time when captured.

Holmes teaches the audio and video originally captured at a different rate from one another and later associated via time when captured (“synchronizes two or more streams of data", col. 1, lines 34-44; "rate of video clock... ratio... adjustable to yield a fixed audio sampling rate", col. 5, lines 33-48;”synch pulses… distinguished by their timing relationships”, col. 4, lines 33-45; where ratios are generally known to be more than just a 1 to 1 ratio, and so the teaching of adjustable ratios implies that the sample rates of the video and audio are not the same; Alternatively, the rates are “different” in nature, since video is, for example, pixel/image frame rate and audio is audio sampling rate, which means that they are different in nature and not in number. The synch pulses

also are an indicator of a time of frame capture because it provides a reference point for comparing when one frame was captured relative to another and where to link the corresponding audio)

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian and Lubiarz, to include the teaching of Holmes of the audio and video originally captured at a different rate from one another and later associated via time when captured in order to prevent either the audio or video from running ahead of the other at any point which could throw off later analysis, as described by Holmes (col. 2, lines 44-64).

17. Regarding **claim 21**, KATSUMI further teaches including instructions for associating other portions of the audio with different ones of the visual features that indicate the mouth is not moving ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

18. Regarding **claim 22**, KATSUMI further teaches instructions for:

identifying second visual features from the video that indicate a different mouth of another speaker is moving or not moving ("images of terminals determined as having speaking attendees", KATSUMI, column 7, lines 56-57); and

associating different portions of the audio with selective ones of the second visual features that indicate the different mouth is moving ("if an audio signal contains a noise such as a page turning noise and voices of other people along with a voice of a conference attendee, since an image of the motion of the lip portion of a conference attendee can be detected from a video signal, the speaking attendee can be determined", KATSUMI, column 7, lines 63-67).

19. Regarding **claim 23**, NEFIAN further teaches instructions for:

processing a neural network ("neural network", NEFIAN, paragraph [0014]) to detect the face of the speaker ("speaker's face in a video sequence", NEFIAN, paragraph [0014]); and

processing a vector matching algorithm to detect movements of the mouth of the speaker within the detected face (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions).

20. Regarding **claim 24**, KATSUMI further teaches that the instructions for associating further include instructions for matching same time slices associated with a time that the portions of the audio were captured and the same time during which the selective ones of the visual features were captured within the video ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents

a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63).

21. Regarding **claim 25**, KATSUMI teaches an apparatus, residing in a computer-accessible medium, comprising:

face detection logic ("detects at least the lip portion of a conference attendee from the video signal", KATSUMI, column 6, lines 23-34);

mouth detection logic ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25); and

audio-video matching logic, wherein the face detection logic detects a face of a speaker within a video ("detects at least the lip portion of a conference attendee", KATSUMI, column 6, lines 23-34), the mouth detection logic detects and monitors movement and non-movement of a mouth included within the face of the video ("extracts the change amount of the shape of the lip portion", KATSUMI, column 6, lines 24-25), and the audio-video matching logic matches captured audio with any movements identified by the mouth detection logic ("when the 'speaking attendee determination information based on audio signal' represents a voice and the 'speaking attendee determination information based on video signal' represents a change of the shape of the lip portion simultaneously", KATSUMI, column 6, lines 58-63)

wherein the video and audio are initially captured together separated for analysis ("speaking attendee determination information based on audio signal", KATSUMI,

column 6, lines 50-51; col. 6, lines 57-67; col. 4, lines 34-46; Figure 3; See Response to Arguments);..

However, KATSUMI does not disclose: specific frequencies occurring within captured audio during a training session and for a same time slice of that training session, and wherein the mouth is detected as moving by changes in pixels that represent the mouth within the face that occur from frame to frame of the video.

In the same field of audiovisual processing, NEFIAN teaches:

specific frequencies occurring within captured audio ("13 MFCC coefficients extracted from a window of 20 ms", NEFIAN, paragraph [0055]) during a training session ("training network and speech recognition module 18", NEFIAN, paragraph [0012]) and for a same time slice of that training session ("discrete nodes at time t for each HMM are conditioned by the discrete nodes at time t1 of all the related HMMs", NEFIAN, paragraph [0023]), and wherein the mouth is detected as moving by changes in pixels that represent the mouth within the face that occur from frame to frame of the video (see NEFIAN, paragraphs [0018]-[0020], a series of vector calculations are performed on the pixels representing the mouth regions)

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use the audiovisual matching method of NEFIAN with the speaker determination system of KATSUMI in order to "improve the performance of speech recognition" (NEFIAN, paragraph [0003]).

Katsumi, in view of Nefian, fail to teach where matching specific frequencies occurring within captured audio is to determine when the speaker is speaking and when the speaker is not speaking.

Lubiarz teaches where matching specific frequencies occurring within captured audio is to determine when the speaker is speaking and when the speaker is not speaking ("speech recognition", col. 2, lines 28-37; "detects silence… detects the presence of voice activity… optimized… for each of the frequency bands", col. 5, lines 17-34; exceeding a threshold indicates speech in a frequency band, and performing the appropriate processing as per Katsumi and Nefian includes matching the speech portion with the appropriate video).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian, to include the teaching of Lubiarz of where matching specific frequencies occurring within captured audio is to determine when the speaker is speaking and when the speaker is not speaking, in order to ensure that the most effective processing for a particular type of signal is applied to that type of signal, as described by Lubiarz (col. 1, lines 5-10).

Katsumi, in view of Nefian and Lubiarz, fail to teach the video and audio are initially captured at a different rate from one another and then re-associated via time when each was captured.

Holmes teaches the video and audio are initially captured at a different rate from one another and then re-associated via time when each was captured ("synchronizes two or more streams of data", col. 1, lines 34-44; "rate of video clock... ratio... adjustable

to yield a fixed audio sampling rate", col. 5, lines 33-48;"synch pulses… distinguished by their timing relationships", col. 4, lines 33-45; where ratios are generally known to be more than just a 1 to 1 ratio, and so the teaching of adjustable ratios implies that the sample rates of the video and audio are not the same; Alternatively, the rates are "different" in nature, since video is, for example, pixel/image frame rate and audio is audio sampling rate, which means that they are different in nature and not in number. The synch pulses also are an indicator of a time of frame capture because it provides a reference point for comparing when one frame was captured relative to another and where to link the corresponding audio)

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Katsumi, in view of Nefian and Lubiarz, to include the teaching of Holmes of the video and audio are initially captured at a different rate from one another and then re-associated via time when each was captured in order to prevent either the audio or video from running ahead of the other at any point which could throw off later analysis, as described by Holmes (col. 2, lines 44-64).

22. Regarding **claim 26**, NEFIAN further teaches that the apparatus is used to configure a Bayesian network which models the speaker speaking ("video data must be fused with audio data using… a coupled hidden Markov model [HMM]", NEFIAN, paragraph [0023], where the HMM is a dynamic Bayesian network).

23. Regarding **claim 27**, NEFIAN further teaches that the face detection logic comprises a neural network ("neural network", NEFIAN, paragraph [0014]).

24. **Claims 7 and 12** are rejected under 35 U.S.C. 103(a) as being unpatentable over Katsumi Patent No.: US 6,369,846 ("KATSUMI") in view of Nefian Pub. No.: US 2003/0212557 ("NEFIAN") and Lubiarz, as applied to Claims 1 and 8, above, and further in view of Van Schyndel Patent No.: US 5,940,118 ("VAN SCHYNDEL").

25. Regarding **claim 7**, the combination of KATSUMI, NEFIAN, and Libuarz teach all the limitations of claim 1.

However, KATSUMI, Nefian, and Lubiarz do not specifically disclose suspending the capturing of audio during periods where select ones of the captured visual features indicate that the speaker is not speaking.

In the same field of audiovisual processing, VAN SCHYNDEL teaches suspending the capturing of audio during periods where select ones of the captured visual features indicate that the speaker is not speaking ("uses optical information to optimally select and/or steer a microphone array in the direction of the talker", VAN SCHYNDEL, column 2, lines 55-58, meaning audio is not captured for someone who is not speaking).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use selectable microphones of VAN SCHYNDEL with the speaker determination system of KATSUMI and audiovisual matching method

of NEFIAN, and Lubiarz, in order to not restrict a talker's movement or position (VAN SCHYNDEL, column 2, lines 60-61).

26. Regarding **claim 12**, the combination of KATSUMI, NEFIAN, and Lubiarz, teach all the limitations of claim 8.

However KATSUMI, NEFIAN, and Lubiarz, do not specifically disclose suspending the capturing of audio when the analysis does not detect the mouths moving for the first and second speakers.

In the same field of audiovisual processing, VAN SCHYNDEL teaches suspending the capturing of audio when the analysis does not detect the mouths moving for the first and second speakers ("uses optical information to optimally select and/or steer a microphone array in the direction of the talker", VAN SCHYNDEL, column 2, lines 55-58, meaning audio is not captured for someone who is not speaking).

Therefore, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to use selectable microphones of VAN SCHYNDEL with the speaker determination system of KATSUMI and audiovisual matching method of NEFIAN, and Lubiarz, in order to not restrict a talker's movement or position (VAN SCHYNDEL, column 2, lines 60-61).

Conclusion

4. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP

§ 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to ERIC YEN whose telephone number is (571)272-4249. The examiner can normally be reached on M-F 7:30-4:00.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Patrick Edouard can be reached on 571-272-7603. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

EY 10/23/08

/Patrick N. Edouard/
Supervisory Patent Examiner, Art Unit 2626